Convergence of Online Mirror Descent Algorithms

Yunwen Lei

Joint work with Professor Ding-Xuan Zhou (City University of Hong Kong).
Outline

1 Background
Outline

1. Background

2. Objectives
Outline

1. Background
2. Objectives
3. Main Results
Outline

1. Background
2. Objectives
3. Main Results
4. Proof
Background
Consider optimization problem

\[
\min_{w \in \mathbb{R}^d} F(w) = \frac{1}{n} \sum_{i=1}^{n} \phi(y_i, \langle w, x_i \rangle) + r(w)
\]

- data fitting term
- regularizer

- examples \( z_t = (x_t, y_t) \) drawn from measure \( \rho \) on \( Z = \mathcal{X} \times \mathcal{Y} \)
- linear model \( x \rightarrow \langle w, x \rangle \), loss function \( \phi : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+ \)
- **big data** era: large sample size \( n \), dimension \( d \)

**Gradient Descent**: with step size \( \{ \eta_t \} \) and initial \( w_1 \in \mathbb{R}^d \)

\[
w_{t+1} = w_t - \eta_t \nabla F(w_t), \quad t \in \mathbb{N}
\]

- **first-order** method: only use information on gradients
- **Hilbert space**: \( w_t \) in **primal** space, \( \nabla F(w_t) \) in **dual** space
- computationally **expensive**: gradient calculation requires going through all examples
Mirror Descent and Interpretation

- A primal space \((\mathcal{W}, \| \cdot \|)\) with its dual \((\mathcal{W}^*, \| \cdot \|_*)\)
- A differentiable **mirror map** \(\Psi : \mathcal{W} \rightarrow \mathbb{R}\), \(\sigma\)-strongly convex

\[
D_{\Psi}(w, \tilde{w}) := \Psi(w) - \left[\Psi(\tilde{w}) + \langle w - \tilde{w}, \nabla \Psi(\tilde{w}) \rangle\right] \geq \frac{\sigma}{2} \| w - \tilde{w} \|^2
\]

first-order approximation of \(\Psi(w)\) at \(\tilde{w}\)

- \(D_{\Psi}(w, \tilde{w})\) called the **Bregman distance** between \(w\) and \(\tilde{w}\)
- with step size \(\{\eta_t\}\) (Nemirovsky and Yudin, 1983)

\[
\nabla \Psi(w_{t+1}) = \nabla \Psi(w_t) - \eta_t \nabla F(w_t)
\]

As a gradient descent in the **dual space** (Nemirovsky and Yudin, 1983)

- \(\nabla \Psi\) maps \(w_t \in \mathcal{W}\) to \(\nabla \Psi(w_t) \in \mathcal{W}^*\)
- performs gradient descent in \(\mathcal{W}^*\) as \(\nabla F(w_t) \in \mathcal{W}^*\)

**use mirror map to capture geometry** of problem by \((\mathcal{W}, \| \cdot \|)\)
Mirror Descent and Interpretation

As a nonlinear subgradient method (Beck and Teboulle, 2003)

\[ w_{t+1} = \arg \min_{w \in \mathcal{W}} \left( F(w_t) + \langle w - w_t, \nabla F(w_t) \rangle + \eta_t^{-1} D_{\Psi}(w, w_t) \right) \]

- first-order approximation of \( F(w) \) at \( w_t \)
- stabilizer

▷ if \( \Psi(w) = \frac{1}{2} \| w \|^2_2 \), \( D_{\Psi}(w, w_t) = \frac{1}{2} \| w - w_t \|^2_2 \), reduce to GD

use mirror map to induce Bregman distance instead of Euclidean distance

Typical choice of \( \Psi \)

▷ \( \Psi(w) = \frac{1}{2} \| w \|^2_p, p \in (1, 2], \) then

\[ (\mathcal{W}, \| \cdot \|) = (\mathbb{R}^d, \| \cdot \|_p), \quad (\mathcal{W}^*, \| \cdot \|_*) = (\mathbb{R}^d, \| \cdot \|_{p^{-1}}) \]

Banach space \((\mathbb{R}^d, \| \cdot \|_p)\) with \( p = 1 + \frac{1}{\log d} \) is preferable in the sparse case, logarithmic dependence on \( d \)
Online Mirror Descent

Motivation

- examples \((x_t, y_t)\) arrives **sequentially** from a measure \(\rho\)
- objective function

\[
F(w) = \mathbb{E}_Z[f(w, Z)], f(w, Z) = \phi(\langle w, X \rangle, Y) + r(w)
\]

Online Mirror Descent

\[
\nabla \Psi(w_{t+1}) = \nabla \Psi(w_t) - \eta_t \nabla_w[f(w_t, z_t)], \quad t \in \mathbb{N}. \tag{1}
\]

- an **instantaneous** regularized loss
  \(f(w, z_t) = \phi(\langle w, x_t \rangle, y_t) + r(w)\) built upon arrival of \(z_t\)
- computationally **cheap**: gradient calculation on an example
- cover **stochastic** setting by uniformly drawing \(z_t\) in a sample
Online Mirror Descent Algorithm—Instantiations

**Online Gradient Descent:** $\Psi = \Psi_2$

$$w_{t+1} = w_t - \eta_t \nabla_w [f(w_t, z_t)].$$

**Randomized Kaczmarz Algorithm:**

$\Psi = \Psi_2, r(w) = 0, \phi(a, y) = \frac{1}{2} (a - y)^2$  

(Lin and Zhou, 2015)

$$w_{t+1} = w_t - \eta_t [\langle w_t, x_t \rangle - y_t] x_t.$$  

**Online $p$-norm Algorithm:** $\Psi = \Psi_p, p \in (1, 2]$  

(Shalev-Shwartz et al., 2012)

$$\begin{cases} v_{t+1} = v_t - \eta_t \nabla_w [f(w_t, z_t)], \\ w_{t+1} = \| v_{t+1} \|_p^{2-p} (\text{sgn}(v_{t+1}(i)) |v_{t+1}(i)|)_{i=1}^d. \end{cases}$$
Objectives
Objectives

This study aims to address these questions:

- What is the role of step sizes in the algorithm? necessary and sufficient conditions for the convergence of \( w_t \) to

\[
    w^* = \arg \min_{w \in \mathcal{W}} F(w)\?
\]

- Can we establish both lower and upper bounds for convergence rates matching up to a constant factor?

- What is the essential difference between online mirror descent and its batch analog?
Main Results
Definitions

A differentiable function $f : \mathcal{W} \to \mathbb{R}$ is $\sigma$-strongly convex w.r.t $\| \cdot \|$ if $D_f(w, \tilde{w}) \geq \frac{\sigma}{2} \| w - \tilde{w} \|^2$, and $L$-strongly smooth w.r.t. $\| \cdot \|$ if $D_f(w, \tilde{w}) \leq \frac{L}{2} \| w - \tilde{w} \|^2$.

Definition

We say $\nabla \Psi$ satisfies an **incremental condition** (of order 1) at infinity if there exists a constant $C_\Psi > 0$ s.t.

$$
\| \nabla \Psi(w) \|_* \leq C_\Psi (1 + \| w \|), \quad \forall w \in \mathcal{W}.
$$

(2)

- intuition: the dual norm of $\nabla \Psi(w)$ is bounded by a **linear** function of $\| w \|$.
- used to show the **necessary condition** for the convergence.
- satisfied by strongly-smooth mirror maps and $p$-norm divergence $\Psi_p$. 
Definitions

Definition

We say the convexity of $\Psi$ is controlled by that of $F$ around $w^*$ with a convex function $\Omega : [0, \infty) \to \mathbb{R}_+$ satisfying $\Omega(0) = 0$ and $\Omega(u) > 0$ for $u > 0$ if the pair $(\Psi, F)$ satisfies

$$\langle w^* - w, \nabla F(w^*) - \nabla F(w) \rangle \geq \Omega (D_\Psi (w^*, w)), \quad \forall w \in \mathcal{W}. \quad (3)$$

- related to strong convexity

$$\langle w^* - w, \nabla F(w^*) - \nabla F(w) \rangle = D_F (w, w^*) + D_F (w^*, w).$$

- typical choices of $\Omega$ include $\Omega(u) = Cu^\alpha$, $\alpha \geq 1$.
  - strongly smooth $\Psi$, strongly convex $F$, (3) holds with $\Omega(u) = C_{\Psi,L}u$ for some $C_{\Psi,L} > 0$.
  - $\Psi = \Psi_p$, strongly convex $F$, (3) holds with $\Omega(u) = C_{\Psi,L}\Omega_p(u)$

$$\Omega_p(u) = \begin{cases} 
  u + \frac{1}{\tau_p} - 1, & \text{if } u \geq 1, \\
  \frac{1}{\tau_p} u^{\tau_p}, & \text{if } 0 \leq u < 1,
\end{cases} \quad \tau_p := \frac{2}{\min\{p, 3 - p\}}. \quad (4)$$
Definitions

\[ \Omega_p(u) \]

\( p = \frac{4}{3} \) (red line), \( p = \frac{3}{2} \) (blue line) and \( p = 2 \) (black line).

\[ \Omega_2 \text{ defined by (4) with } p = 2 \text{ is a Huber loss!} \] (Huber et al., 1964)
Main Results—Positive Variances

Assumptions

- positive variances: \( \inf_{w \in \mathcal{W}} \mathbb{E}_Z [\| \nabla_w f(w, Z) \|_*] > 0 \)
- smoothness: \( f(\cdot, z) \) is \( L \)-strongly smooth for a.e. \( z \in Z \)
- \( \nabla \Psi \) continuous at \( w^* \), satisfies incremental condition at \( \infty \)
- pair \( (\Psi, F) \) meets (3) at \( w^* \) with convex \( \Omega : [0, \infty) \to \mathbb{R}_+ \)

Results: \( \lim_{t \to \infty} \mathbb{E}[D_{\Psi}(w^*, w_t)] = 0 \) if and only if

\[
\lim_{t \to \infty} \eta_t = 0 \text{ and } \sum_{t=1}^{\infty} \eta_t = \infty
\]

Furthermore:

- If \( \Psi \) is strongly smooth and \( \lim_{t \to \infty} \eta_t = 0 \), then
  \[
  \mathbb{E}[D_{\Psi}(w^*, w_T)] \geq \frac{\tilde{C}}{T - t_0 + 1}, \quad \forall T \geq t_0
  \]
- If \( \Omega(u) = \sigma_F u \) and \( \eta_t = \frac{4}{(t+1)\sigma_F} \), then \( \mathbb{E}[D_{\Psi}(w^*, w_T)] = O \left( \frac{1}{T} \right) \).
Main Results—Zero Variances

Assumptions

- zero variances: \( \mathbb{E}_Z [\| \nabla_w[f(w^*, Z)] \|_*] = 0 \)
- smoothness: \( f(\cdot, z) \) is \( L \)-strongly smooth for a.e. \( z \in Z \)
- \( \nabla \Psi \) continuous at \( w^* \), satisfies incremental condition at \( \infty \)
- pair \((\Psi, F)\) meets (3) at \( w^* \) with convex \( \Omega : [0, \infty) \rightarrow \mathbb{R}_+ \)
- \( w_1 \neq w^* \), \( \eta_t \leq \frac{\sigma \Psi}{(2 + \kappa)L} \) for some \( \kappa > 0 \)

Results: \( \lim_{t \to \infty} \mathbb{E}[D_{\Psi}(w^*, w_t)] = 0 \) if and only if \( \sum_{t=1}^{\infty} \eta_t = \infty \).

Furthermore:

- If \( \Omega(u) = \sigma_F u \) and \( \eta_t \equiv \eta_1 < \frac{\sigma \Psi}{2L} \), then

\[
\left(1 - \frac{2L\eta_1}{\sigma \Psi}\right)^T D_{\Psi}(w^*, w_1) \leq \mathbb{E}[D_{\Psi}(w^*, w_T)] \leq \left(1 - \frac{\sigma_F \eta_1}{2}\right)^T D_{\Psi}(w^*, w_1).
\]

for cases with zero variances, online mirror descent behaves analogously to mirror descent!
Main Results—Almost Sure Convergence

Assumptions

- smoothness: $f(\cdot, z)$ is $L$-strongly smooth for a.e. $z \in Z$
- $\nabla \Psi$ continuous at $w^*$, satisfies incremental condition at $\infty$
- pair $(\Psi, F)$ meets (3) at $w^*$ with convex $\Omega : [0, \infty) \to \mathbb{R}_+$
- step size sequence satisfies

$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty$$

Results: $\{\|w_t - w^*\|^2\}_{t \in \mathbb{N}}$ converges to 0 almost surely
Main Results—Specific Applications

Assumptions—regularization scheme

- \( R := \sup_{x \in \mathcal{X}} \|x\|_* < \infty, \| \cdot \| = \| \cdot \|_2 \)
- the loss function \( \phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+ \) is \( \ell_\phi \)-strongly smooth
- regularized loss \( f(w, z) = \phi(\langle w, x \rangle, y) + \lambda \|w\|_2^2 \) with \( \lambda > 0 \)
- \( \Psi \) is either a \( p \)-norm divergence \( \Psi = \Psi_p \) with \( 1 < p \leq 2 \) or a strongly smooth mirror map

Strongly smooth loss functions:

- least square: \( \phi(y, a) = (y - a)^2 \)
- logistic loss: \( \phi(y, a) = \log(1 + \exp(-ya)) \)
- 2-norm hinge loss: \( \phi(y, a) = \max(0, 1 - ya)^2 \)
- \( \phi(a, y) = 1/(1 + e^{ay}) \)
Main Results—Specific Applications

Results:

(a) Assume $\inf_{w \in \mathcal{W}} \mathbb{E}_Z [\| \nabla w [f(w, Z)] \|_*] > 0$. Then
\[
\lim_{t \to \infty} \mathbb{E} [\| w_t - w^* \|^2] = 0 \text{ if and only if } \lim_{t \to \infty} \eta_t = 0 \text{ and } \sum_{t=1}^{\infty} \eta_t = \infty. \text{ Furthermore, if } \Psi \text{ strongly smooth, then for some } \tilde{T}_1, \tilde{C} > 0 \text{ s.t. } \mathbb{E} [\| w_T - w^* \|^2] \geq \tilde{C} T^{-1} \text{ for } T \geq \tilde{T}_1. \text{ If } \eta_t = \frac{4}{(t+1)\sigma} \text{ for some } \sigma > 0, \text{ then } \mathbb{E} [\| w_T - w^* \|^2] = O(T^{-1}).$

(b) If $\mathbb{E}_Z [\| \nabla w [f(w^*, Z)] \|_*] = 0$ and for some $\kappa > 0$, $\eta_t \leq \frac{\sigma \Psi}{2(\ell \phi R^2 + \lambda)(2+\kappa)}$. Then $\lim_{t \to \infty} \mathbb{E} [\| w_t - w^* \|^2] = 0$ if and only if $\sum_{t=1}^{\infty} \eta_t = \infty$. Furthermore, if $\Psi$ is strongly smooth and $\eta_t \equiv \eta_1 < \frac{\sigma \Psi}{4(\ell \phi R^2 + \lambda)}$, then there exist $\tilde{c}_1, \tilde{c}_2 \in (0, 1)$ s.t.
\[
\tilde{c}_1 T \| w_1 - w^* \|^2 \leq \mathbb{E} [\| w_T - w^* \|^2] \leq \tilde{c}_2 T \| w_1 - w^* \|^2, \quad \forall T \in \mathbb{N}.
\]

(c) If the step size sequence satisfies $\sum_{t=1}^{\infty} \eta_t = \infty$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$, then $\lim_{t \to \infty} \| w_t - w^* \| = 0$ almost surely.
Main Results—Specific Applications

Assumptions—no regularization

- \( R := \sup_{x \in \mathcal{X}} \|x\|_* < \infty \)
- unregularized least squares: \( f(w, z) = \frac{1}{2} (y - \langle w, x \rangle)^2 \)
- \( \Psi \) is either a \( p \)-norm divergence \( \Psi = \Psi_p \) with \( 1 < p \leq 2 \) or a strongly smooth mirror map
- \( \nabla \Psi(w_1) \) belonging to the range of \( C_X^T, C_X := \mathbb{E}_Z[XX^T] \)
- Define \( w_\rho = \min_{w \in \mathcal{W}} \{ \Psi(w) : C_X w = \mathbb{E}_Z[XY] \} \).

- strongly smooth mirror map
- randomized Kaczmarz algorithm (Lin and Zhou, 2015)

\[ \Psi(w) = \frac{1}{2} \|w\|_2^2. \]

- smoothed linearized Bregman iteration (Cai et al., 2009)

\[ \Psi^{(\epsilon, \lambda)}(w) = \lambda \sum_{i=1}^{d} g_\epsilon(w(i)) + \frac{1}{2} \|w\|_2^2, \]

where \( g_\epsilon(\xi) := \frac{\xi^2}{2\epsilon} \) for \( |\xi| \leq \epsilon \) and \( |\xi| - \frac{\epsilon}{2} \) for \( |\xi| > \epsilon \).
Main Results—Specific Applications

Results:

(a) Assume \( \inf_{w \in \mathcal{W}} [ |Y - \langle w, X \rangle| \|X\|_* ] > 0 \). Then
\[
\lim_{t \to \infty} \mathbb{E}[\|w_t - w_\rho\|^2] = 0 \text{ if and only if } \lim_{t \to \infty} \eta_t = 0 \text{ and } \\
\sum_{t=1}^\infty \eta_t = \infty.
\]
Furthermore, if \( \Psi \) strongly smooth, then for some \( \tilde{T}_1, \tilde{C} > 0 \) s.t. \( \mathbb{E}[\|w_T - w_\rho\|^2] \geq \tilde{C}T^{-1} \) for \( T \geq \tilde{T}_1 \). If \( \eta_t = \frac{4}{(t+1)\sigma} \) for some \( \sigma > 0 \), then \( \mathbb{E}[\|w_T - w_\rho\|^2] = O(T^{-1}) \).

(b) If \( \mathbb{E}_Z [ |Y - \langle w, X \rangle| \|X\|_* ] = 0 \) and for some \( \kappa > 0 \),
\[
\eta_t \leq \frac{\sigma\psi}{(2+\kappa)R^2}.
\]
Then \( \lim_{t \to \infty} \mathbb{E}[\|w_t - w_\rho\|^2] = 0 \) if and only if \( \sum_{t=1}^\infty \eta_t = \infty \). Furthermore, if \( \Psi \) is strongly smooth and
\( \eta_t \equiv \eta_1 < \frac{\sigma\psi}{(2+\kappa)R^2} \), then there exist \( \tilde{c}_1, \tilde{c}_2 \in (0, 1) \) s.t.
\[
\tilde{c}_1^T \|w_1 - w_\rho\|^2 \leq \mathbb{E}[\|w_T - w_\rho\|^2] \leq \tilde{c}_2^T \|w_1 - w_\rho\|^2, \quad \forall T \in \mathbb{N}.
\]

(c) If the step size sequence satisfies \( \sum_{t=1}^\infty \eta_t = \infty \) and \( \sum_{t=1}^\infty \eta_t^2 < \infty \), then \( \lim_{t \to \infty} \|w_t - w_\rho\| = 0 \) almost surely.
Discussions

Existing studies consider convergence of online gradient descent algorithms

**Euclidean space:**
almost sure convergence studied under assumptions (Bottou, 1998)

\[
\inf_{\|w - w^*\|_2^2 > \epsilon} \langle w - w^*, \nabla F(w) \rangle > 0, \quad \forall \epsilon > 0, \quad \|\nabla F(w)\|_2^2 \leq A + B\|w - w^*\|_2^2, \quad \forall w \in \mathcal{W}
\]

**Reproducing kernel Hilbert space:**
sufficient conditions established for regression, classification (Smale and Yao, 2006; Ying and Zhou, 2006)

**Randomized Kaczmarz Algorithm:** (Lin and Zhou, 2015)

- sufficient and necessary conditions established
- analysis only applies to least squares loss and \(\Psi = \Psi_2\)
- require restrictions \(0 < \eta_t < 2\)
- lower bounds \(\|w_t - w^*\|_2^2 \geq \tilde{C}t^{-2}\) not tight
Proof
A key Identity

**One-step progress** of OMD in terms of the excess Bregman distance \( D_\Psi(w^*, w_{t+1}) - D_\Psi(w^*, w_t) \)

**Lemma**

*The following identity holds for* \( t \in \mathbb{N} \)

\[
\mathbb{E}_{z_t}[D_\Psi(w^*, w_{t+1})] - D_\Psi(w^*, w_t) = \eta_t \langle w^* - w_t, \nabla F(w_t) \rangle + \mathbb{E}_{z_t}[D_\Psi(w_t, w_{t+1})]. \tag{5}
\]

**Idea of Analysis**: control \( \mathbb{E}[D_\Psi(w^*, w_{t+1})] \) from both **above and lower** in terms of \( \mathbb{E}[D_\Psi(w^*, w_t)] \), using strong smooth of \( F \), strong convexity of \( \Psi \) and convexity of pair \((\Psi, F)\)
Positive Variances—**Necessary Conditions**

**necessary condition:** $\lim_{t \to \infty} \eta_t = 0$.

Denote $\sigma := \inf_{w \in \mathcal{W}} \mathbb{E}_Z [||\nabla_w f(w, Z)||_*]$.

- with incremental condition and continuity of $\Psi$ at $w^*$, we show

  $$\lim_{t \to \infty} \mathbb{E} [D\Psi(w^*, w_t)||_*] = 0 \implies \lim_{t \to \infty} \mathbb{E} [||\nabla \Psi(w_t) - \nabla \Psi(w^*)||_*] = 0$$

- $\lim_{t \to \infty} \eta_t = 0$ then follows by

  $$\eta_t \sigma \leq \eta_t \mathbb{E}_{z_t} [||\nabla_w f(w_t, z_t)||_*] \leq ||\nabla \Psi(w_t) - \nabla \Psi(w^*)||_* + \mathbb{E}_{z_t} [||\nabla \Psi(w_{t+1}) - \nabla \Psi(w^*)||_*]$$
Positive Variances—Necessary Conditions

**necessary condition**: $\sum_{t=1}^{\infty} \eta_t = \infty$.

- by $L_F$-strong smoothness of $F$ and $\sigma_\Psi$-strong convexity of $\Psi$, we get

$$\langle w^* - w_t, \nabla F(w_t) \rangle \geq -L_F \|w^* - w_t\|^2 \geq -\frac{2L_F}{\sigma_\Psi} D_\Psi(w^*, w_t).$$

- this plugged into (5) gives ($a = 2L_F\sigma_\Psi^{-1}$)

$$\mathbb{E}[D_\Psi(w^*, w_{t+1})] \geq (1 - a\eta_t) \mathbb{E}[D_\Psi(w^*, w_t)] + \mathbb{E}[D_\Psi(w_t, w_{t+1})].$$

(6)

- apply this inequality repeatedly gives

$$\mathbb{E}[D_\Psi(w^*, w_{T+1})] \geq \exp \left(-2a \sum_{t=t_0+1}^{T} \eta_t \right) \mathbb{E}[D_\Psi(w^*, w_{t_0+1})].$$
Positive Variances—**Sufficient Conditions**

by $D_g(w, \tilde{w}) = D^*_g(\nabla g(\tilde{w}), \nabla g(w))$ ($g^*$ is Fenchel-conjugate)

$$D_\Psi(w_t, w_{t+1}) = D^*_\Psi(\nabla \Psi(w_{t+1}), \nabla \Psi(w_t)) \leq \frac{1}{2\sigma_\Psi} \| \nabla \Psi(w_{t+1}) - \nabla \Psi(w_t) \|_2^2$$

$$= \frac{\eta_t^2}{2\sigma_\Psi} \| \nabla_w [f(w_t, z_t)] \|_2^2.$$ 

by $L$-strong smoothness of $f(\cdot, z)$, we derive (co-coercivity)

$$\| \nabla_w [f(w_t, z_t)] \|_2^2 \leq 2\| \nabla_w [f(w_t, z_t)] - \nabla_w [f(w^*, z_t)] \|_2^2 + 2\| \nabla_w [f(w^*, z_t)] \|_2^2$$

$$\leq 2L \langle w^* - w_t, \nabla F(w^*) - \nabla F(w_t) \rangle + 2\| \nabla_w [f(w^*, z_t)] \|_2^2$$

plugged into one-step progress identity (5) gives

$$\mathbb{E}_{z_t}[D_\Psi(w^*, w_{t+1})] \leq D_\Psi(w^*, w_t) - \frac{\eta_t}{2} \langle w^* - w_t, \nabla F(w^*) - \nabla F(w_t) \rangle + \frac{\eta_t^2}{\sigma_\Psi} \mathbb{E}_{z_t} \left[ \| \nabla_w [f(w^*, z_t)] \|_2^2 \right]$$

$$\leq D_\Psi(w^*, w_t) - \frac{\eta_t}{2} \Omega(D_\Psi(w^*, w_t)) + b\eta_t^2, \quad b := \frac{1}{\sigma_\Psi} \mathbb{E}_Z \left[ \| \nabla_w [f(w^*, Z)] \|_2^2 \right]$$

**convexity** of $\Omega$ further implies

$$A_{t+1} \leq A_t - \frac{\eta_t}{2} \Omega(A_t) + b\eta_t^2, \quad A_t := \mathbb{E}[D_\Psi(w^*, w_t)]$$

convergence of $A_t$ follows by $\lim_{t \to \infty} \eta_t = 0$ and $\sum_{t=1}^{\infty} \eta_t = \infty$. 

**Convergence of Online Mirror Descent Algorithms**


